Finite Mixture Models

GEOFFREY McLACHLAN

DAVID PEEL

Department of Mathematics
The University of Queensland
Contents

Preface xix

1 General Introduction 1
  1.1 Introduction 1
    1.1.1 Flexible Method of Modeling 1
    1.1.2 Initial Approach to Mixture Analysis 2
    1.1.3 Impact of EM Algorithm 3
  1.2 Overview of Book 4
  1.3 Basic Definition 6
  1.4 Interpretation of Mixture Models 7
  1.5 Shapes of Some Univariate Normal Mixtures 9
    1.5.1 Mixtures of Two Normal Homoscedastic Components 9
    1.5.2 Mixtures of Univariate Normal Heteroscedastic Components 11
  1.6 Modeling of Asymmetrical Data 14
  1.7 Normal Scale Mixture Model 17
  1.8 Spurious Clusters 17
  1.9 Incomplete-Data Structure of Mixture Problem 19
  1.10 Sampling Designs for Classified Data 21
  1.11 Parametric Formulation of Mixture Model 22
## CONTENTS

1.12 Nonparametric ML Estimation of a Mixing Distribution 23
1.13 Estimation of Mixture Distributions 24
1.14 Identifiability of Mixture Distributions 26
1.15 Clustering of Data via Mixture Models 29
   1.15.1 Mixture Likelihood Approach to Clustering 29
   1.15.2 Decision-Theoretic Approach 30
   1.15.3 Clustering of I.I.D. Data 31
   1.15.4 Image Segmentation or Restoration 32
1.16 Hidden Markov Models 33
1.17 Testing for the Number of Components in Mixture Models 34
1.18 Brief History of Finite Mixture Models 35
1.19 Notation 37

2 ML Fitting of Mixture Models 40
   2.1 Introduction 40
   2.2 ML Estimation 40
   2.3 Information Matrices 41
   2.4 Asymptotic Covariance Matrix of MLE 42
   2.5 Properties of MLEs for Mixture Models 42
   2.6 Choice of Root 44
   2.7 Test for a Consistent Root 44
      2.7.1 Basis of Test 44
      2.7.2 Example 2.1: Likelihood Function with Two Maximizers 45
      2.7.3 Formulation of Test Statistic 45
   2.8 Application of EM Algorithm for Mixture Models 47
      2.8.1 Direct Approach 47
      2.8.2 Formulation as an Incomplete-Data Problem 48
      2.8.3 E-Step 49
      2.8.4 M-Step 49
      2.8.5 Assessing the Implied Error Rates 50
   2.9 Fitting Mixtures of Mixtures 51
   2.10 Maximum a Posteriori Estimation 52
   2.11 An Aitken Acceleration-Based Stopping Criterion 52
   2.12 Starting Values for EM Algorithm 54
      2.12.1 Specification of an Initial Parameter Value 54
      2.12.2 Random Starting Values 55
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.12.3</td>
<td>Example 2.2: Synthetic Data Set 1</td>
<td>56</td>
</tr>
<tr>
<td>2.12.4</td>
<td>Deterministic Annealing EM Algorithm</td>
<td>57</td>
</tr>
<tr>
<td>2.13</td>
<td>Stochastic EM Algorithm</td>
<td>61</td>
</tr>
<tr>
<td>2.14</td>
<td>Rate of Convergence of the EM Algorithm</td>
<td>61</td>
</tr>
<tr>
<td>2.14.1</td>
<td>Rate Matrix for Linear Convergence</td>
<td>61</td>
</tr>
<tr>
<td>2.14.2</td>
<td>Rate Matrix in Terms of Information Matrices</td>
<td>62</td>
</tr>
<tr>
<td>2.15</td>
<td>Information Matrix for Mixture Models</td>
<td>63</td>
</tr>
<tr>
<td>2.15.1</td>
<td>Direct Evaluation of Observed Information Matrix</td>
<td>63</td>
</tr>
<tr>
<td>2.15.2</td>
<td>Extraction of Observed Information Matrix in Terms of the Complete-Data Log Likelihood</td>
<td>64</td>
</tr>
<tr>
<td>2.15.3</td>
<td>Approximations to Observed Information Matrix: I.I.D. Case</td>
<td>64</td>
</tr>
<tr>
<td>2.15.4</td>
<td>Supplemented EM Algorithm</td>
<td>66</td>
</tr>
<tr>
<td>2.15.5</td>
<td>Conditional Bootstrap Approach</td>
<td>67</td>
</tr>
<tr>
<td>2.16</td>
<td>Provision of Standard Errors</td>
<td>68</td>
</tr>
<tr>
<td>2.16.1</td>
<td>Information-Based Methods</td>
<td>68</td>
</tr>
<tr>
<td>2.16.2</td>
<td>Bootstrap Approach to Standard Error Approximation</td>
<td>68</td>
</tr>
<tr>
<td>2.17</td>
<td>Speeding up Convergence</td>
<td>70</td>
</tr>
<tr>
<td>2.17.1</td>
<td>Introduction</td>
<td>70</td>
</tr>
<tr>
<td>2.17.2</td>
<td>Louis' Method</td>
<td>71</td>
</tr>
<tr>
<td>2.17.3</td>
<td>Quasi-Newton Methods</td>
<td>72</td>
</tr>
<tr>
<td>2.17.4</td>
<td>Hybrid Methods</td>
<td>72</td>
</tr>
<tr>
<td>2.18</td>
<td>Outlier Detection from a Mixture</td>
<td>74</td>
</tr>
<tr>
<td>2.18.1</td>
<td>Introduction</td>
<td>74</td>
</tr>
<tr>
<td>2.18.2</td>
<td>Modified Likelihood Ratio Test</td>
<td>74</td>
</tr>
<tr>
<td>2.19</td>
<td>Partial Classification</td>
<td>75</td>
</tr>
<tr>
<td>2.20</td>
<td>Partial Nonrandom Classification</td>
<td>76</td>
</tr>
<tr>
<td>2.20.1</td>
<td>Introduction</td>
<td>76</td>
</tr>
<tr>
<td>2.20.2</td>
<td>A Nonrandom Model</td>
<td>77</td>
</tr>
<tr>
<td>2.20.3</td>
<td>Asymptotic Relative Efficiencies</td>
<td>77</td>
</tr>
<tr>
<td>2.21</td>
<td>Classification ML Approach</td>
<td>79</td>
</tr>
<tr>
<td>3</td>
<td>Multivariate Normal Mixtures</td>
<td>81</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>81</td>
</tr>
<tr>
<td>3.2</td>
<td>Heteroscedastic Components</td>
<td>81</td>
</tr>
<tr>
<td>3.3</td>
<td>Homoscedastic Components</td>
<td>83</td>
</tr>
<tr>
<td>3.4</td>
<td>Standard Errors</td>
<td>83</td>
</tr>
</tbody>
</table>
4.3 Conjugate Priors 119
4.4 Markov Chain Monte Carlo 120
  4.4.1 Posterior Simulation 120
  4.4.2 Perfect Sampling 121
4.5 Exponential Family Components 121
4.6 Normal Components 122
  4.6.1 Conjugate Priors 122
  4.6.2 Gibbs Sampler 123
4.7 Prior on Number of Components 124
4.8 Noninformative Settings 125
  4.8.1 Improper Priors 125
  4.8.2 Data-Dependent Priors 126
  4.8.3 Markov Prior on Component Means 126
  4.8.4 Reparameterization for Univariate Normal Components 127
4.9 Label-Switching Problem 129
4.10 Prior Feedback Approach to ML Estimation 132
4.11 Variational Approach to Bayesian Estimation 132
4.12 Minimum Message Length 133

5 Mixtures with Nonnormal Components 135
  5.1 Introduction 135
  5.2 Mixed Continuous and Categorical Variables 136
    5.2.1 Location Model-Based Approach 137
    5.2.2 Implementation of Location Model 138
  5.3 Example 5.1: Prostate Cancer Data 139
    5.3.1 Description of Data Set 139
    5.3.2 Fitting Strategy under MULTIMIX 140
  5.4 Generalized Linear Model 142
    5.4.1 Definition 142
    5.4.2 ML Estimation for a Single GLM Component 143
    5.4.3 Quasi-Likelihood Approach 144
  5.5 Mixtures of GLMs 145
    5.5.1 Specification of Mixture Model 145
    5.5.2 ML Estimation via the EM Algorithm 146
    5.5.3 M-Step 147
    5.5.4 Multicycle ECM Algorithm 148
    5.5.5 Choice of the Number of Components 148
  5.6 A General ML Analysis of Overdispersion in a GLM 149
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td>Approaches for Assessing Mixture Order</td>
<td>184</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Main Approaches</td>
<td>184</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Nonparametric Methods</td>
<td>184</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Method of Moments</td>
<td>185</td>
</tr>
<tr>
<td>6.4</td>
<td>Likelihood Ratio Test Statistic</td>
<td>185</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Introduction</td>
<td>185</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Example 6.3: Breakdown in Regularity Conditions</td>
<td>186</td>
</tr>
<tr>
<td>6.5</td>
<td>Distributional Results for the LRTS</td>
<td>187</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Some Theoretical Results</td>
<td>187</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Some Simulation Results</td>
<td>189</td>
</tr>
<tr>
<td>6.5.3</td>
<td>Mixtures of Two Unrestricted Normal Components</td>
<td>190</td>
</tr>
<tr>
<td>6.5.4</td>
<td>Mixtures of Two Exponentials</td>
<td>191</td>
</tr>
<tr>
<td>6.6</td>
<td>Bootstrapping the LRTS</td>
<td>192</td>
</tr>
<tr>
<td>6.6.1</td>
<td>Implementation</td>
<td>192</td>
</tr>
<tr>
<td>6.6.2</td>
<td>Application to Three Real Data Sets</td>
<td>194</td>
</tr>
<tr>
<td>6.6.3</td>
<td>Applications in Astronomy</td>
<td>196</td>
</tr>
<tr>
<td>6.7</td>
<td>Effect of Estimates on $P$-Values of Bootstrapped LRTS</td>
<td>198</td>
</tr>
<tr>
<td>6.7.1</td>
<td>Some Simulation Results</td>
<td>198</td>
</tr>
<tr>
<td>6.7.2</td>
<td>Double Bootstrapping</td>
<td>200</td>
</tr>
<tr>
<td>6.8</td>
<td>Information Criteria in Model Selection</td>
<td>202</td>
</tr>
<tr>
<td>6.8.1</td>
<td>Bias Correction of the Log Likelihood</td>
<td>202</td>
</tr>
<tr>
<td>6.8.2</td>
<td>Akaike’s Information Criterion</td>
<td>203</td>
</tr>
<tr>
<td>6.8.3</td>
<td>Bootstrap-Based Information Criterion</td>
<td>203</td>
</tr>
<tr>
<td>6.8.4</td>
<td>Cross-Validation-Based Information Criterion</td>
<td>205</td>
</tr>
<tr>
<td>6.8.5</td>
<td>Minimum Information Ratio Criterion</td>
<td>206</td>
</tr>
<tr>
<td>6.8.6</td>
<td>Informational Complexity Criterion</td>
<td>207</td>
</tr>
<tr>
<td>6.9</td>
<td>Bayesian-Based Information Criteria</td>
<td>207</td>
</tr>
<tr>
<td>6.9.1</td>
<td>Bayesian Approach</td>
<td>207</td>
</tr>
<tr>
<td>6.9.2</td>
<td>Laplace’s Method of Approximation</td>
<td>208</td>
</tr>
<tr>
<td>6.9.3</td>
<td>Bayesian Information Criterion</td>
<td>209</td>
</tr>
<tr>
<td>6.9.4</td>
<td>Laplace–Metropolis Criterion</td>
<td>210</td>
</tr>
<tr>
<td>6.9.5</td>
<td>Laplace–Empirical Criterion</td>
<td>211</td>
</tr>
<tr>
<td>6.9.6</td>
<td>Reversible Jump Method</td>
<td>212</td>
</tr>
<tr>
<td>6.9.7</td>
<td>MML Principle</td>
<td>212</td>
</tr>
<tr>
<td>6.10</td>
<td>Classification-Based Information Criteria</td>
<td>212</td>
</tr>
<tr>
<td>6.10.1</td>
<td>Classification Likelihood Criterion</td>
<td>212</td>
</tr>
<tr>
<td>6.10.2</td>
<td>Normalized Entropy Criterion</td>
<td>214</td>
</tr>
</tbody>
</table>
6.10.3 Integrated Classification Likelihood Criterion 215
6.11 An Empirical Comparison of Some Criteria 217
6.11.1 Simulated Set 1 218
6.11.2 Simulated Set 2 218
6.11.3 Simulated Set 3 219
6.11.4 Conclusions from Simulations 220

7 Multivariate t Mixtures 221
7.1 Introduction 221
7.2 Previous Work 222
7.3 Robust Clustering 222
7.4 Multivariate t Distribution 223
7.5 ML Estimation of Mixture of t Distributions 224
7.5.1 Application of EM Algorithm 224
7.5.2 E-Step 225
7.5.3 M-Step 227
7.5.4 Application of ECM Algorithm 229
7.6 Previous Work on M-Estimation of Mixture Components 230
7.7 Example 7.1: Simulated Noisy Data Set 231
7.8 Example 7.2: Crab Data Set 234
7.9 Example 7.3: Old Faithful Geyser Data Set 236

8 Mixtures of Factor Analyzers 238
8.1 Introduction 238
8.2 Principal Component Analysis 239
8.3 Single-Factor Analysis Model 240
8.4 EM Algorithm for a Single-Factor Analyzer 241
8.5 Data Visualization in Latent Space 243
8.6 Mixtures of Factor Analyzers 244
8.7 AECM Algorithm for Fitting Mixtures of Factor Analyzers 245
8.7.1 AECM Framework 245
8.7.2 First Cycle 245
8.7.3 Second Cycle 246
8.7.4 Representation of Original Data 248
8.8 Link of Factor Analysis with Probabilistic PCA 248
8.9 Mixtures of Probabilistic PCAs 250
8.10 Initialization of AECM Algorithm 250
8.11 Example 8.1: Simulated Data 252
8.12 Example 8.2: Wine Data 254

9 Fitting Mixture Models to Binned Data 257
9.1 Introduction 257
9.2 Binned and Truncated Data 258
9.3 Application of EM Algorithm 259
  9.3.1 Missing Data 259
  9.3.2 E-Step 260
  9.3.3 M-Step 261
  9.3.4 M-Step for Normal Components 261
9.4 Practical Implementation of EM Algorithm 262
  9.4.1 Computational Issues 262
  9.4.2 Numerical Integration at Each EM Iteration 262
  9.4.3 Integration over Truncated Regions 263
  9.4.4 EM Algorithm for Binned Multivariate Data 264
9.5 Simulations 264
9.6 Example 9.1: Red Blood Cell Data 265

10 Mixture Models for Failure-Time Data 268
10.1 Introduction 268
10.2 Competing Risks 269
  10.2.1 Mixtures of Survival Functions 269
  10.2.2 Latent Failure-Time Approach 270
  10.2.3 ML Estimation for Mixtures of Survival Functions 271
10.3 Example 10.1: Heart-Valve Data 272
  10.3.1 Description of Problem 272
  10.3.2 Mixture Models with Unconstrained Components 273
  10.3.3 Constrained Mixture Models 274
  10.3.4 Conditional Probability of a Reoperation 276
  10.3.5 Advantages of Mixture Model-Based Approach 276
10.4 Long-Term Survivor Model 277
  10.4.1 Definition 277
  10.4.2 Modified Long-Term Survivor Model 278
  10.4.3 Partial ML Approach for Modified Long-Term Survival Model 279
10.4.4 Interpretation of Cure Rate in Presence of Competing Risks 280
10.4.5 Example 9.2: Breast Cancer Data 280
10.5 Analysis of Masked System-Life Data 283
10.5.1 Masked Cause of Failure 283
10.5.2 Application of EM Algorithm 283
10.5.3 Exponential Components 284
10.5.4 Weibull Components 285

11 Mixture Analysis of Directional Data 287
11.1 Introduction 287
11.2 Joint Sets 287
11.3 Directional Data 291
11.4 Initial Work on Clustering of Directional Data 292
11.5 Mixture of Kent Distributions 292
11.6 Moment Estimation of Kent Distribution 293
11.7 Uniform Component for Background Noise 295
11.8 Application of EM Algorithm 296
11.9 Example 11.1: Two Mining Samples 297
11.10 Determining the Number of Joint Sets 298
11.11 Discussion 301

12 Variants of the EM Algorithm for Large Databases 302
12.1 Introduction 302
12.2 Incremental EM Algorithm 303
12.2.1 Introduction 303
12.2.2 Definition of Partial E-Step 303
12.2.3 Block Updating of Sufficient Statistics 303
12.2.4 Justification of IEM Algorithm 305
12.2.5 Gain in Convergence Time 305
12.2.6 IEM Algorithm for Singleton Blocks 306
12.2.7 Efficient Updating Formulas 306
12.3 Simulations for IEM Algorithm 307
12.3.1 Simulation 1 307
12.3.2 Simulation 2 309
12.4 Lazy EM Algorithm 310
12.5 Sparse EM Algorithm 311
12.6 Sparse IEM Algorithm 312
12.6.1 Some Simulation Results 312
12.6.2 Summary of Results for the IEM and SPIEM Algorithms 315

12.7 A Scalable EM Algorithm 316
12.7.1 Introduction 316
12.7.2 Primary Compression of the Data 316
12.7.3 Updating of Parameter Estimates 318
12.7.4 Merging of Sufficient Statistics 319
12.7.5 Secondary Data Compression 319
12.7.6 Tuning Constants 320
12.7.7 Simulation Results 321

12.8 Multiresolution KD-Trees 323
12.8.1 Introduction 323
12.8.2 EM Algorithm Based on Multiresolution KD-Trees 323

13 Hidden Markov Models 326
13.1 Introduction 326
13.2 Hidden Markov Chain 328
13.2.1 Definition 328
13.2.2 Some Examples 329

13.3 Applying EM Algorithm to Hidden Markov Chain Model 329
13.3.1 EM Framework 329
13.3.2 E-Step 330
13.3.3 Forward–Backward Recursions on E-Step 330
13.3.4 M-Step 332
13.3.5 Numerical Instabilities 332

13.4 Hidden Markov Random Field 332
13.4.1 Specification of Markov Random Field 333
13.4.2 Application of EM Algorithm 333
13.4.3 Restoration Step 334
13.4.4 An Improved Approximation to EM Solution 335
13.4.5 Approximate M-Step for Normal Components 336

13.5 Example 13.1: Segmentation of MR Images 336
13.6 Bayesian Approach 338

13.7 Examples of Gibbs Sampling with Hidden Markov Chains 339

Appendix Mixture Software 343
| CONTENTS |
|-----------------|-------|
| A.1 EMMIX       | 343   |
| A.2 Some Other Mixture Software | 345   |
| References      | 349   |
| Author Index    | 395   |
| Subject Index   | 407   |